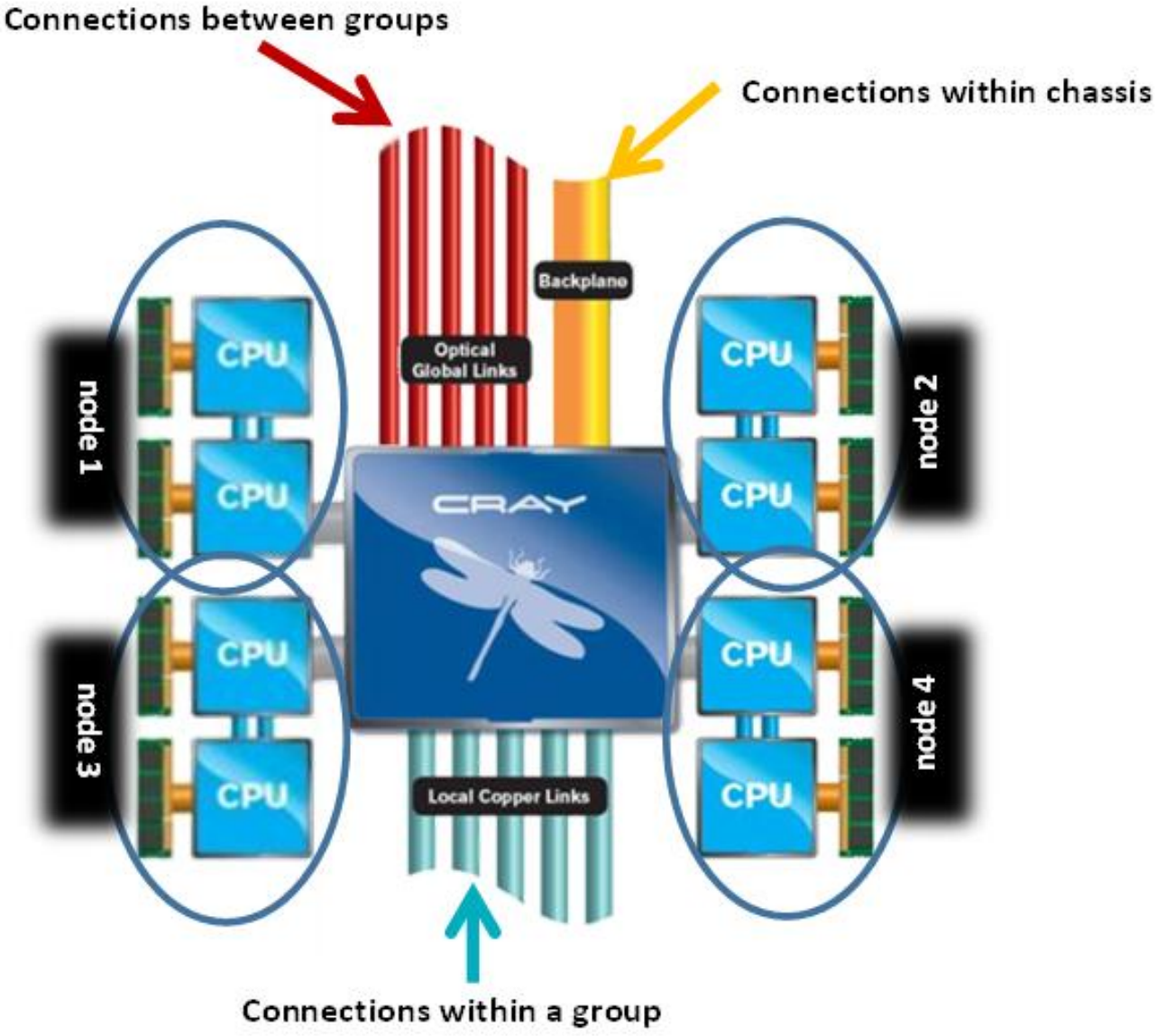
Лекция 13
Управление ресурсами вычислительных систем

Ефимов Александр Владимирович
E-mail: alexandr.v.efimov@sibguti.ru

Курс «Архитектура вычислительных систем»
СибГУТИ, 2020

Архитектура вычислительного модуля Cray XC50



Интерконнект Cray XC50

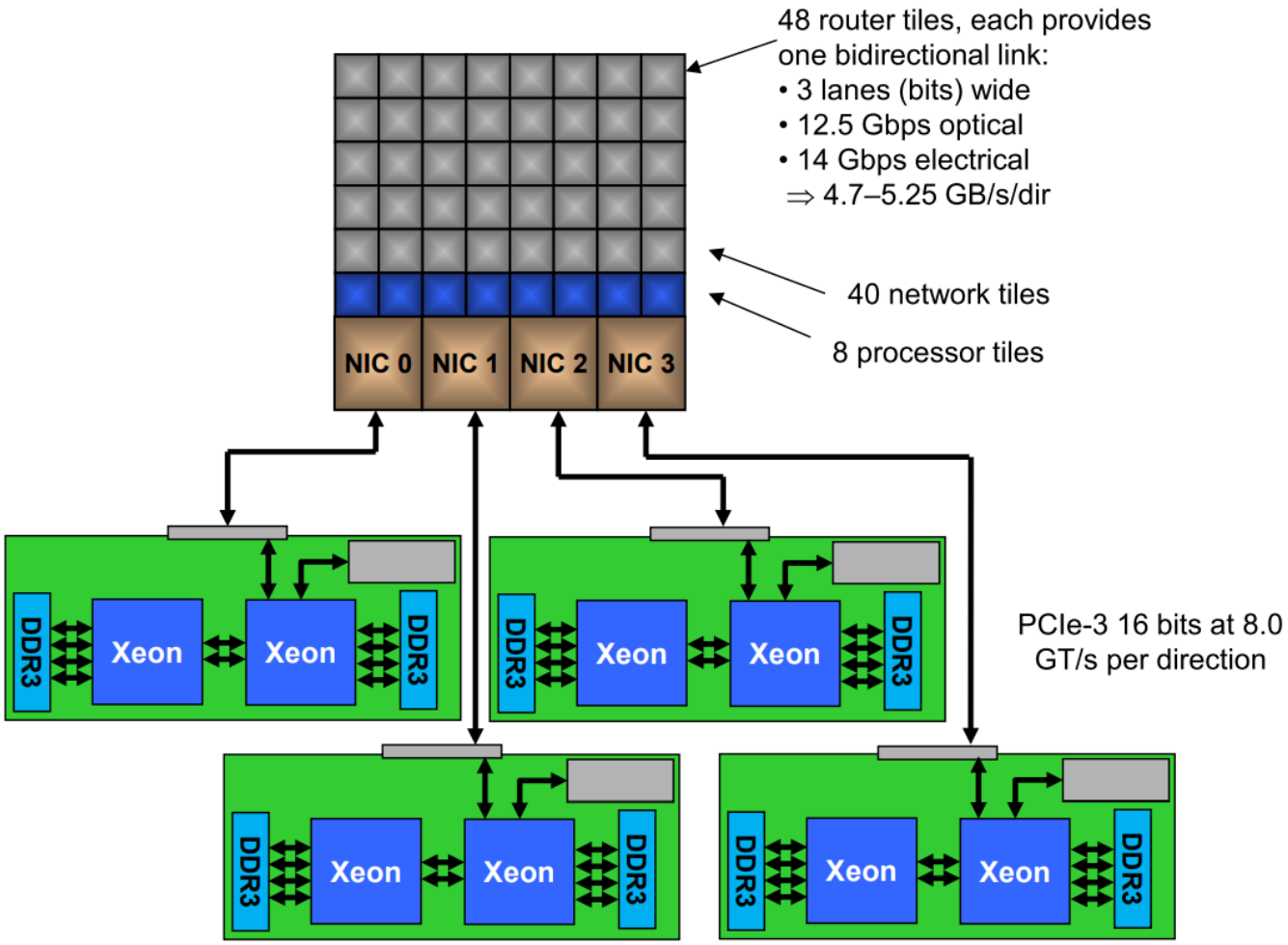
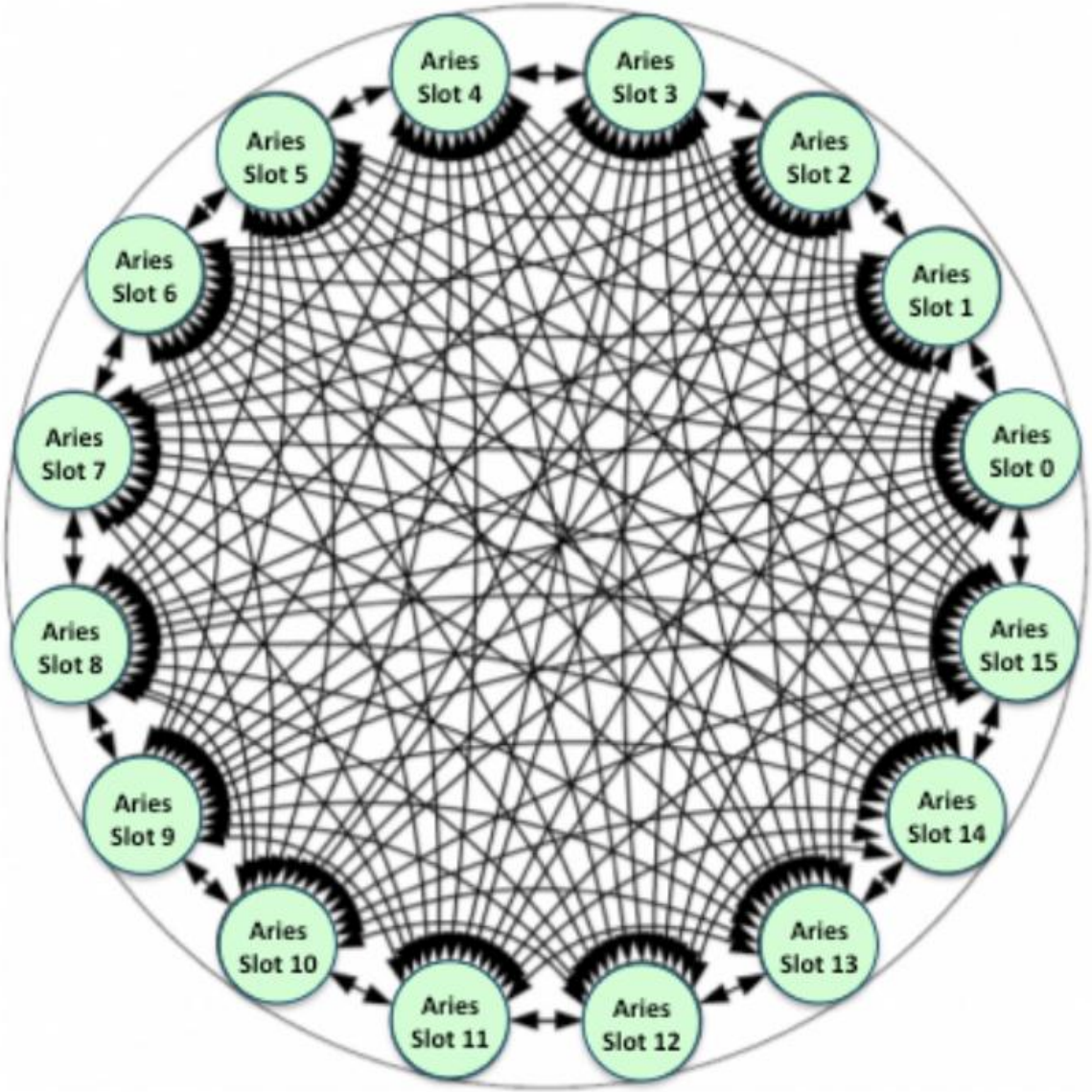
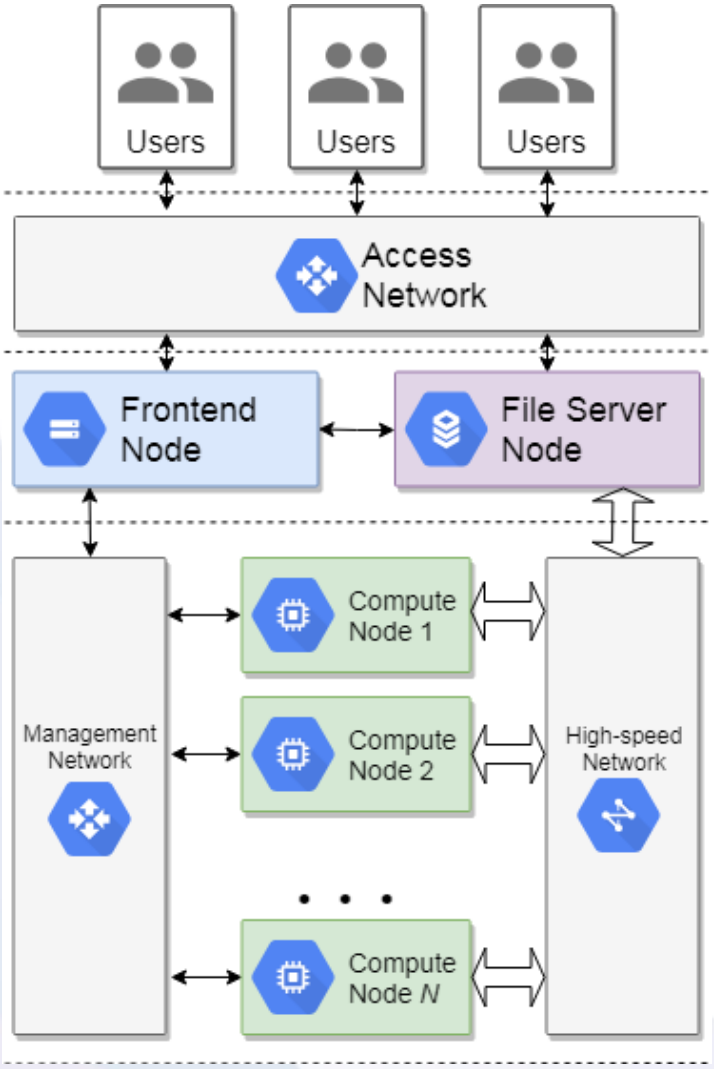


Figure 3: *A single Aries system-on-a-chip device provides network connectivity for the four nodes on a Cray XC blade.*

Топология макроструктуры Dragonfly Cray XC50



Типовая архитектура вычислительного кластера

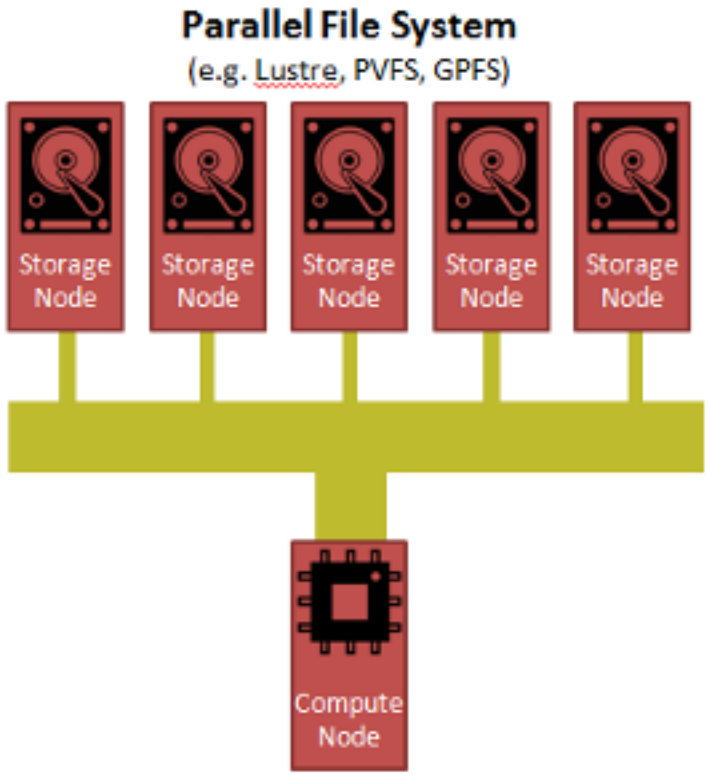


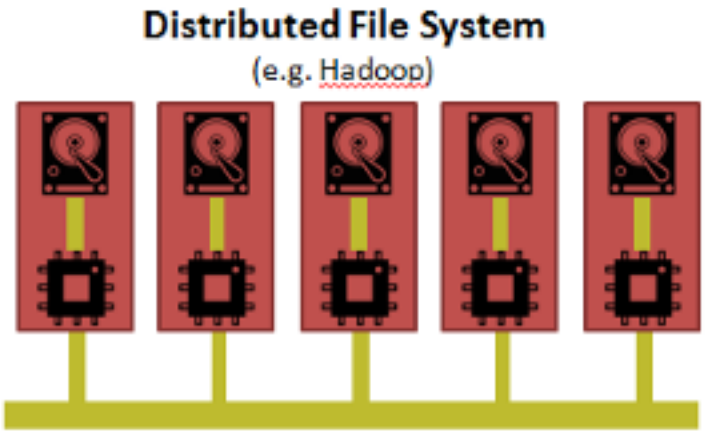
Производительность:
$10^0 – 10^2$ PFlops

Количество вычислительных узлов:
$10^3 – 10^5$ штук

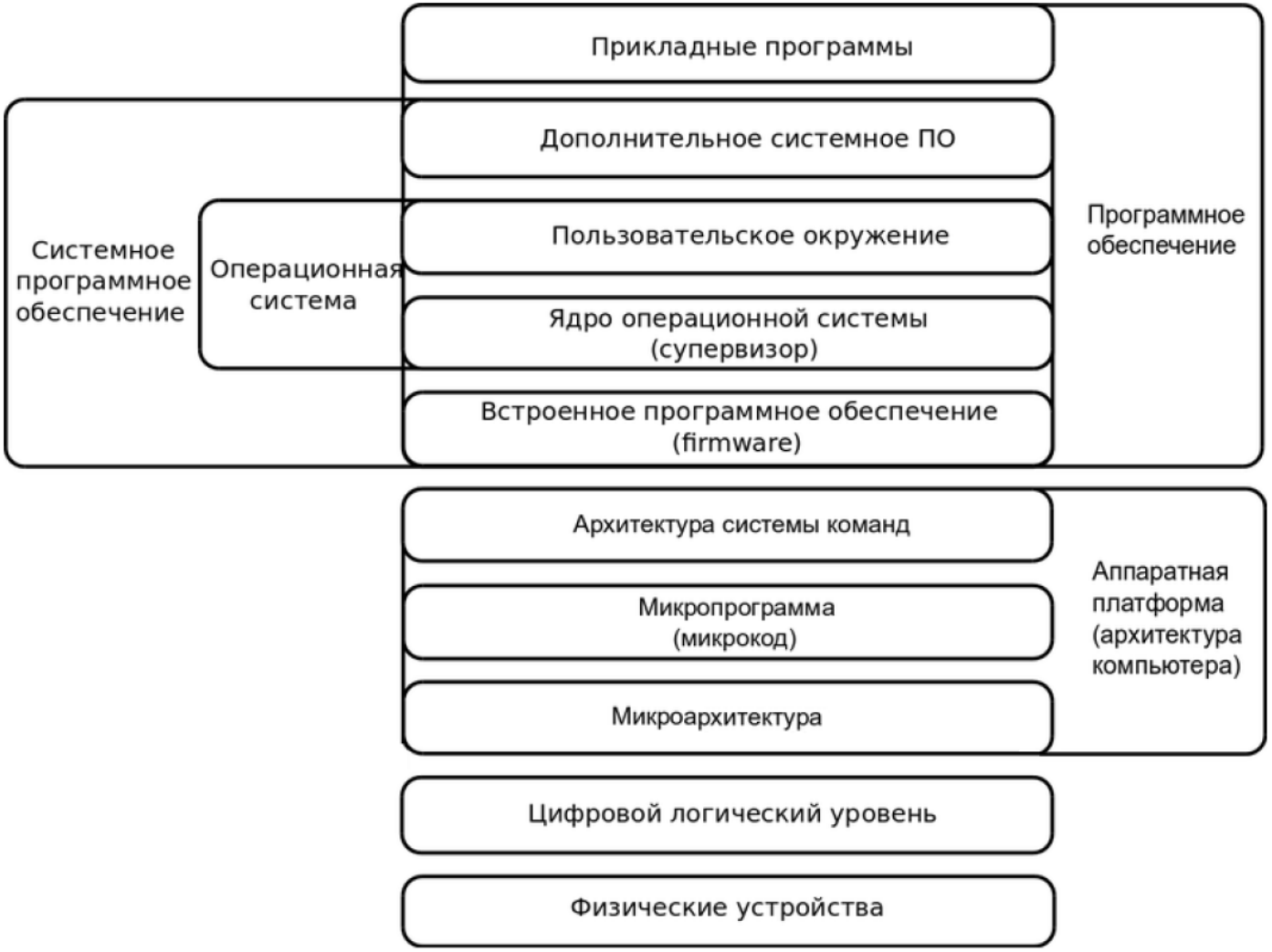
Время наработки на отказ:
$10^0 – 10^2$ часов

Файловые системы





Уровни вычислительного средства

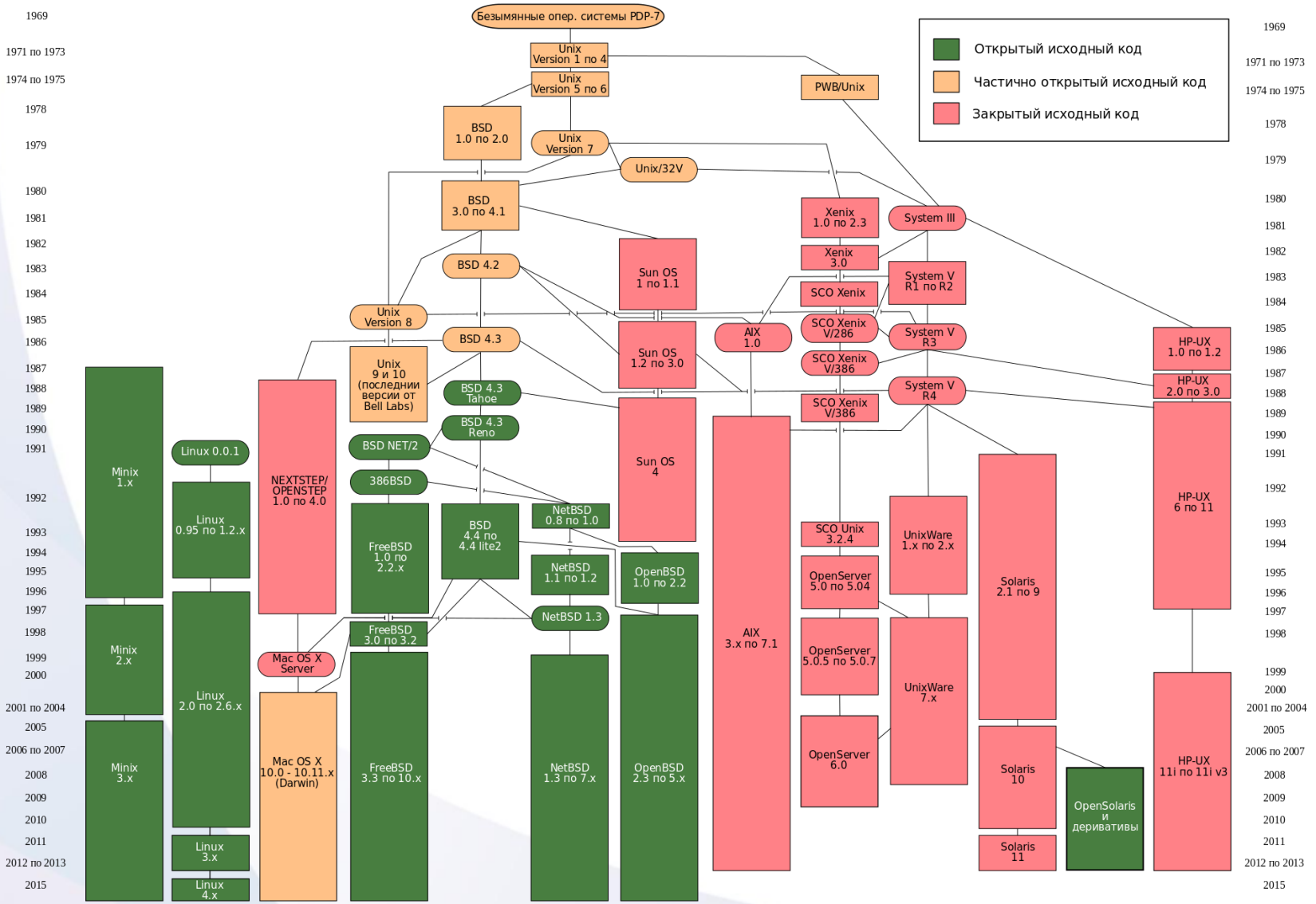


Распределенная операционная система

Распределённая ОС существует как единая операционная система в масштабах вычислительной системы.

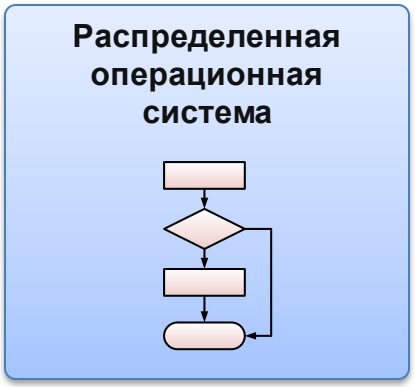
Распределённая ОС, динамически и автоматически распределяет работы по различным машинам системы для параллельной обработки.

Семейство ОС UNIX



Режимы функционирования ВС

Поток параллельных задач

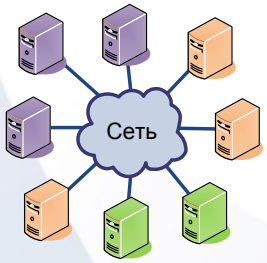


Монозадачный режим

Мультизадачные режимы

Обслуживание потоков задач

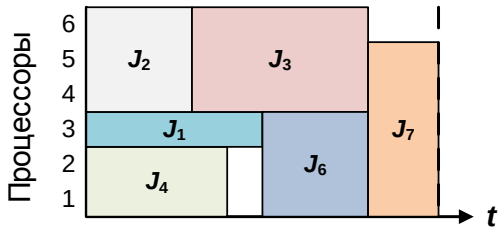
Генерация подсистем в пределах ВС



- Техника теории игр
- Стохастическое программирование

Обработка наборов задач

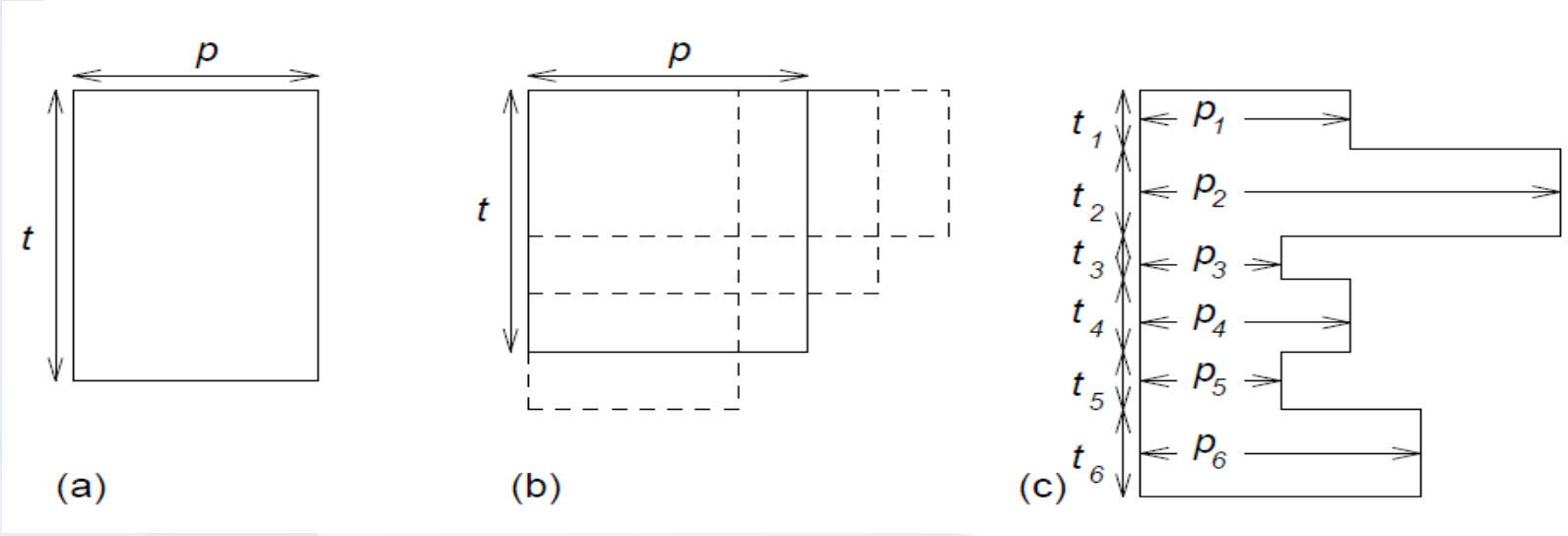
Формирование расписаний решения параллельных задач



Точные, эвристические и стохастические методы и алгоритмы для задач с фиксированными рангами

Классификация задач

Кто определяет число?	**Когда определяют число?**	
	до начала решения	в процессе решения
Пользователь	Жесткая (фиксированная) rigid	изменяющаяся evolving
СУР	масштабируемая moldable	уступчивая malleable



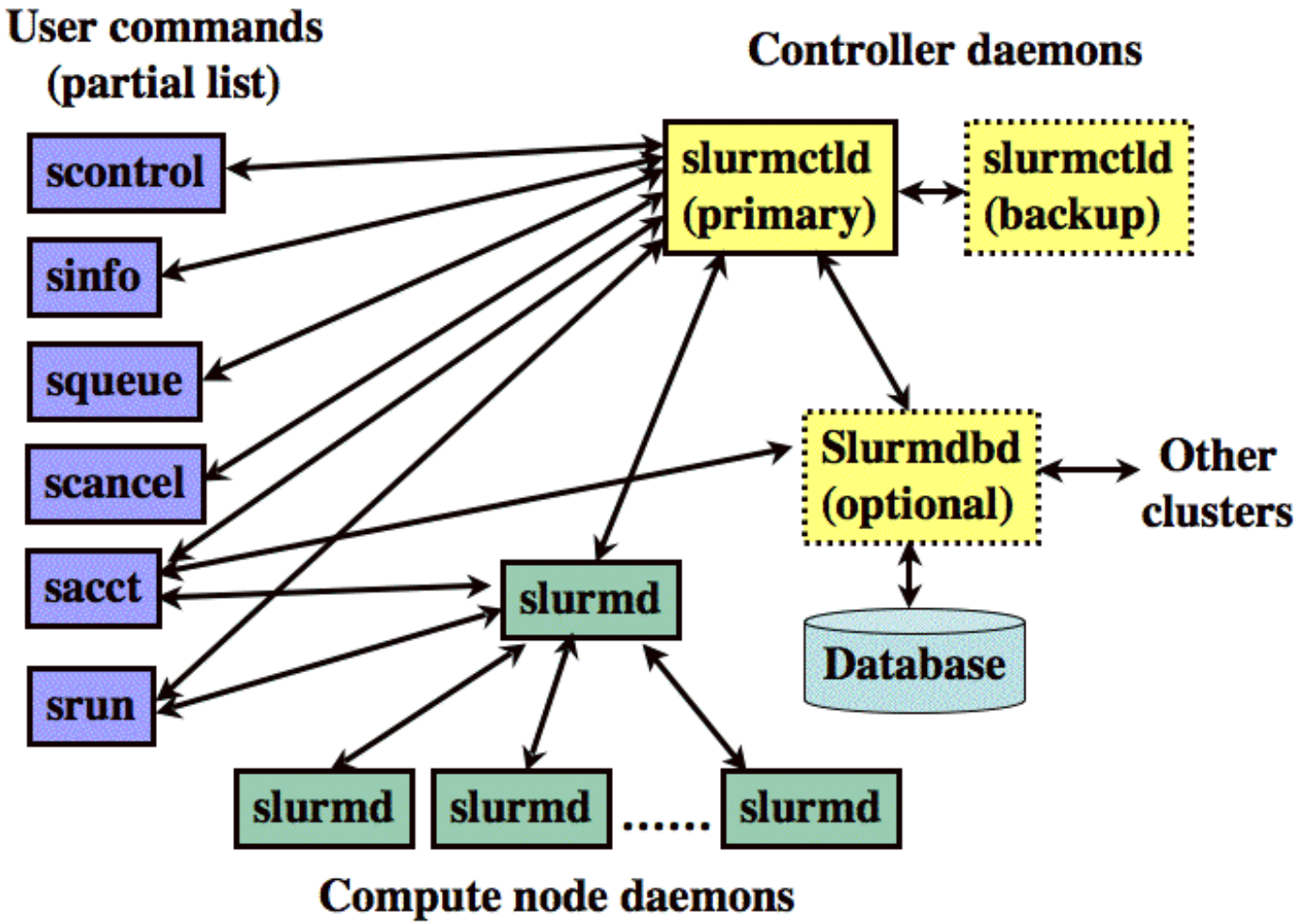
Проблемы при решении задач на ВС

- Распараллелить программу
- Организовать отказоустойчивое выполнение
- Эффективно вложить задачу в структуру системы, т.е. расположить ветви параллельной программы по вычислителям так, чтобы взаимодействие между ними занимало минимум времени.

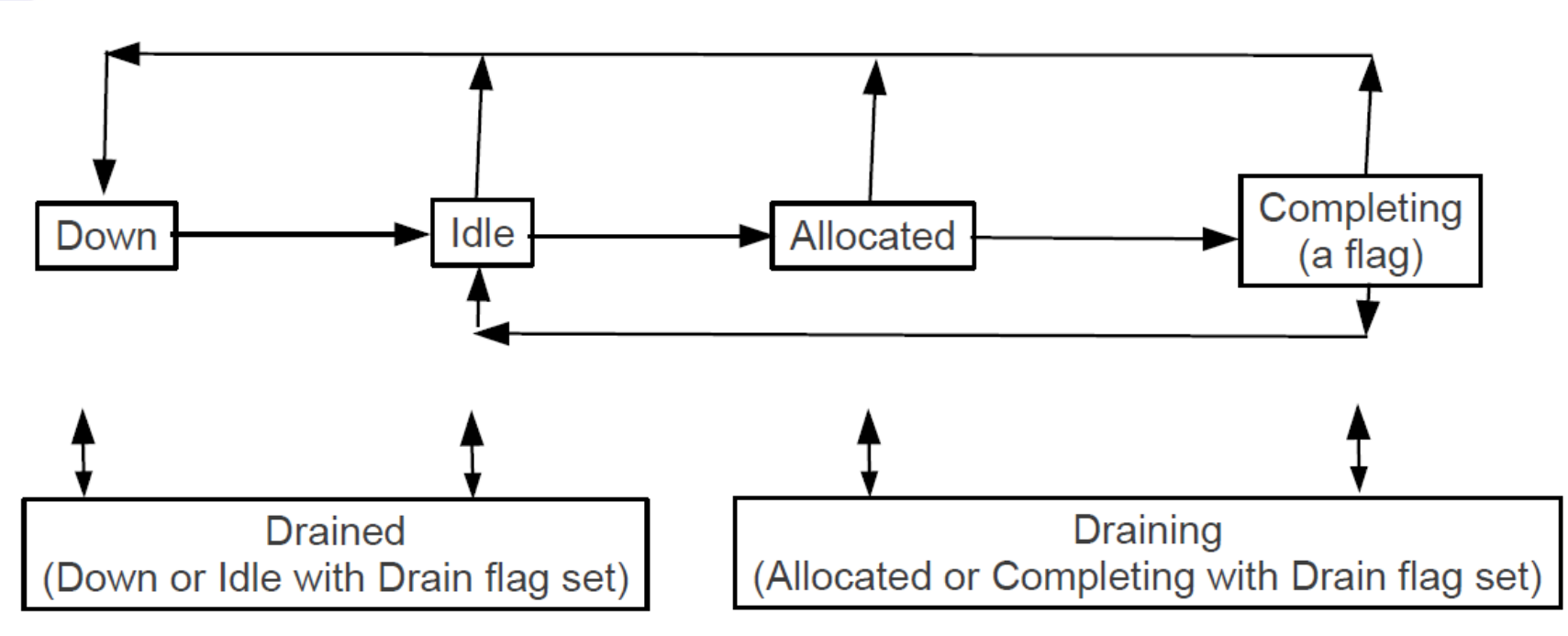
```
$ cat test.job
    #PBS -N Job_Name
    #PBS -q Batch_Name
    #PBS -l nodes=1:ppn=6

    mpiexec ./mpiprog
```

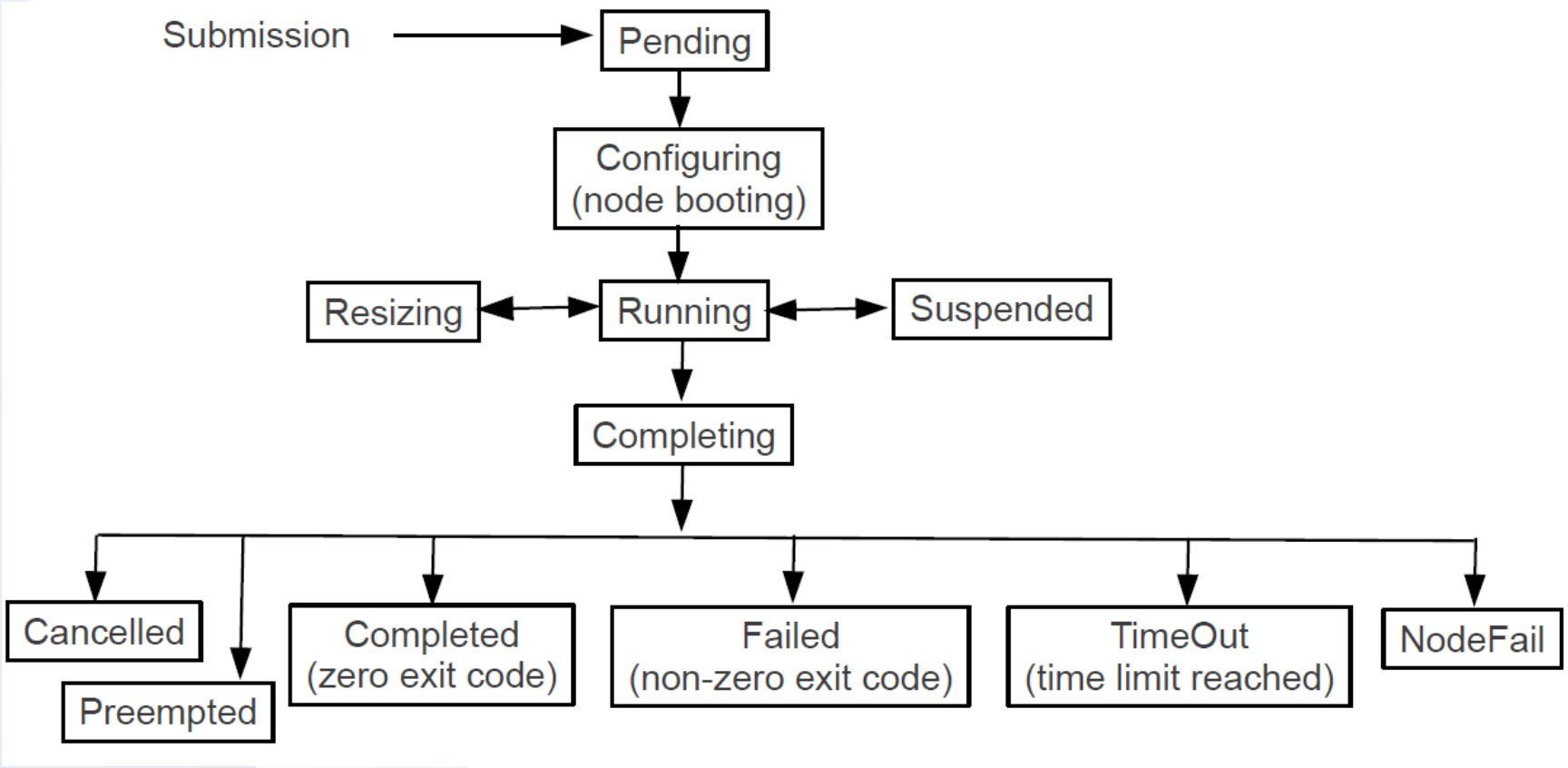
Система управления ресурсами SLURM



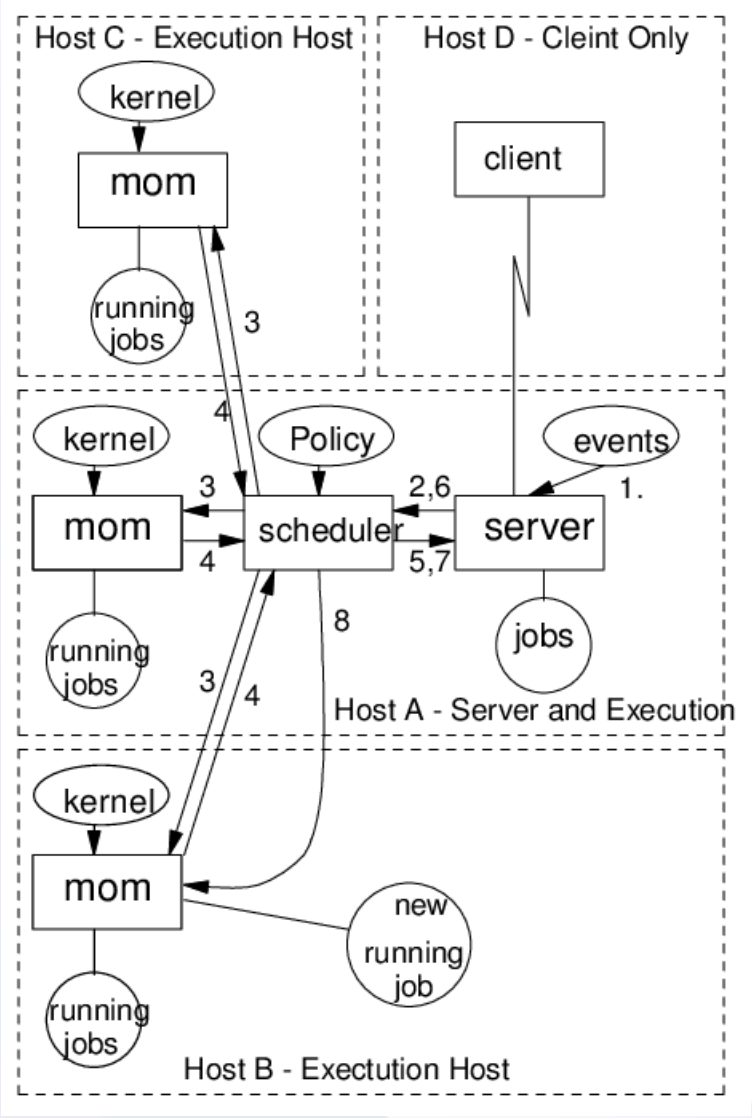
Состояния вычислительного узла



Состояния задачи



Алгоритм планирования



1. Наступление события планирования.
2. Сервер отправляет команду планировщику.

3-4. Планировщик запрашивает информацию о состоянии ресурсов.

5-6. Планировщик запрашивает информацию о задачах.

/* Планирование */

/* First Come First Served (FCFS)*/

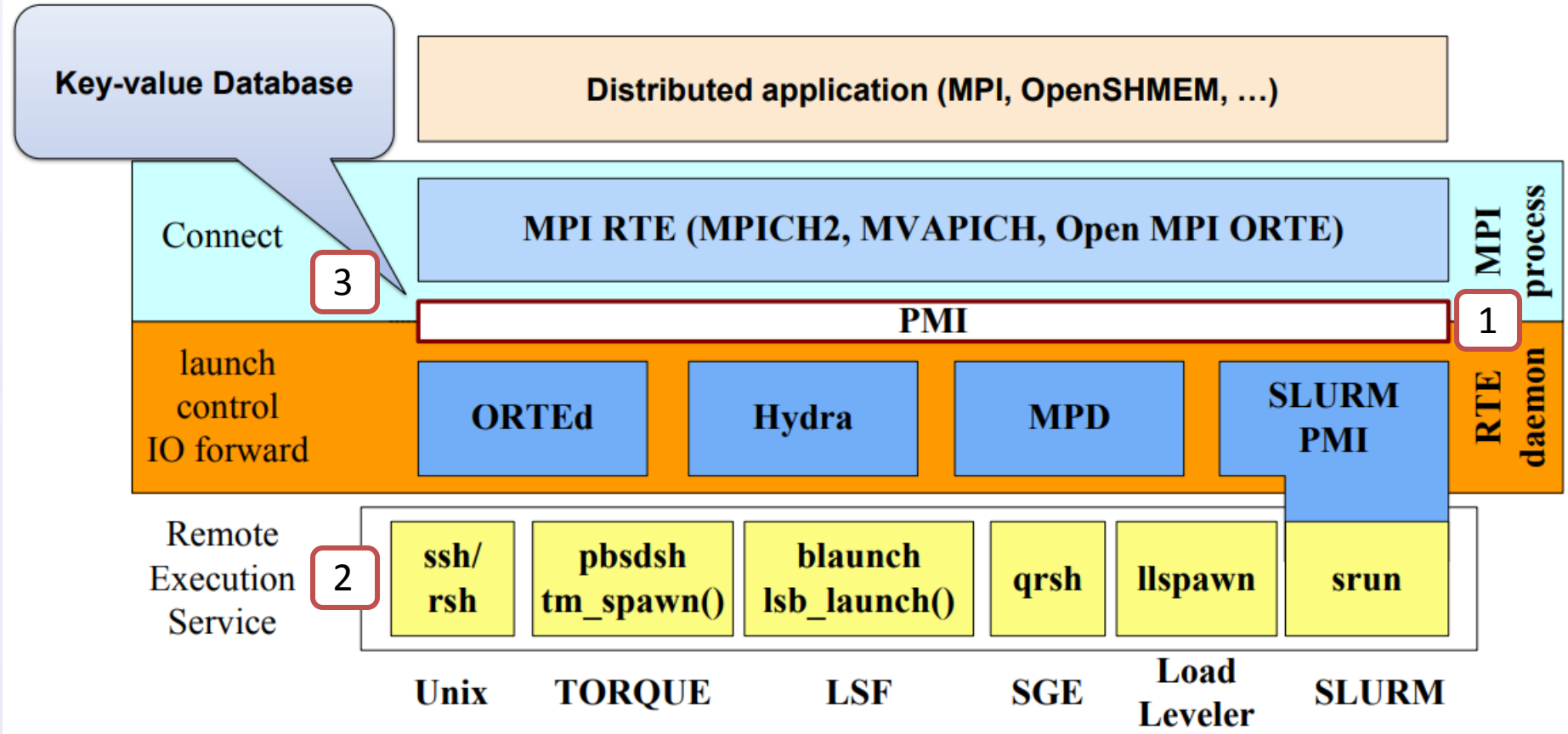
/* Backfilling */

7. Планировщик отправляет запрос на решение задачи серверу.
8. Отправка задачи на вычислительные узлы.

Популярные модели параллельного программирования

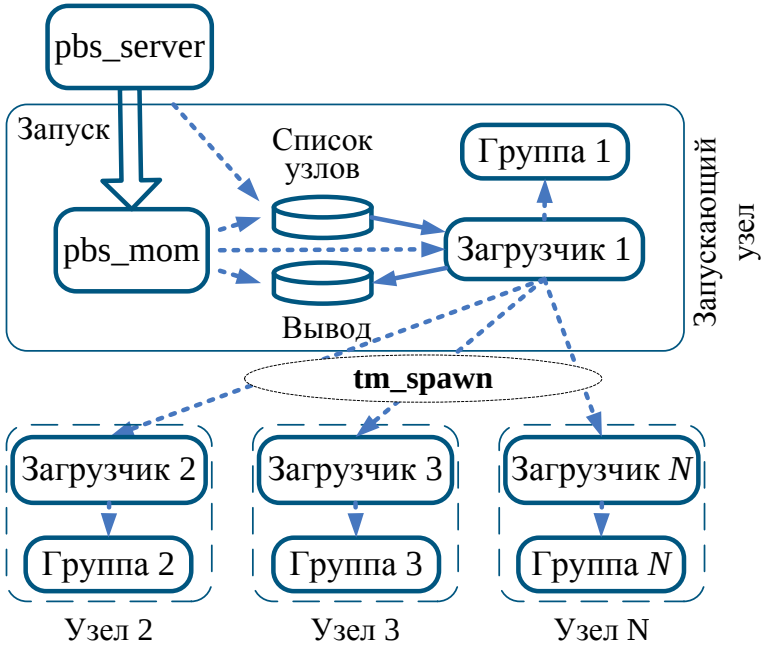
- Модель передачи сообщений (Message Passing Interface, MPI). Библиотеки MPICH, OpenMPI и др.
- Модель разделённого глобального адресного пространство (*partitioned global address space* PGAS). Вся память параллельного вычислительного комплекса (глобальная память) является адресуемой и разделена на логические разделы, каждый из которых локален для процесса или потока. Языки программирования Unified Parallel C, Chapel и X10. Библиотеки PGAS: GASNet и SHMEM.

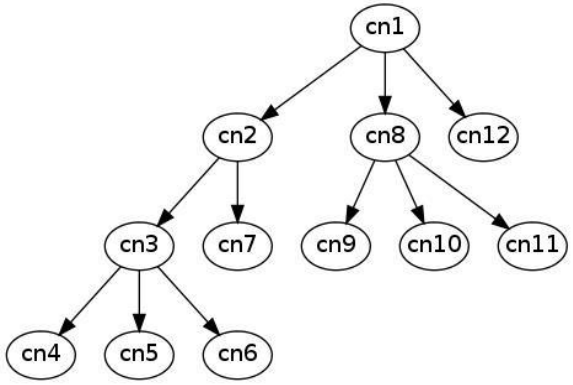
Стек программного обеспечения среды исполнения (RunTime Environment, RTE)

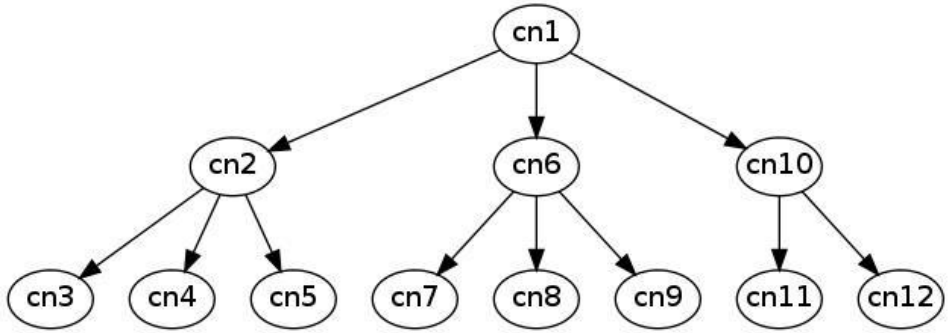


Artem Y. Polyakov, Joshua S. Ladd, Boris I. Karasev Towards Exascale: Leveraging InfiniBand to accelerate the performance and scalability of Slurm jobstart. SuperComputing 2017 (SC'17), USA, Colorado, November 12-17, 2017.

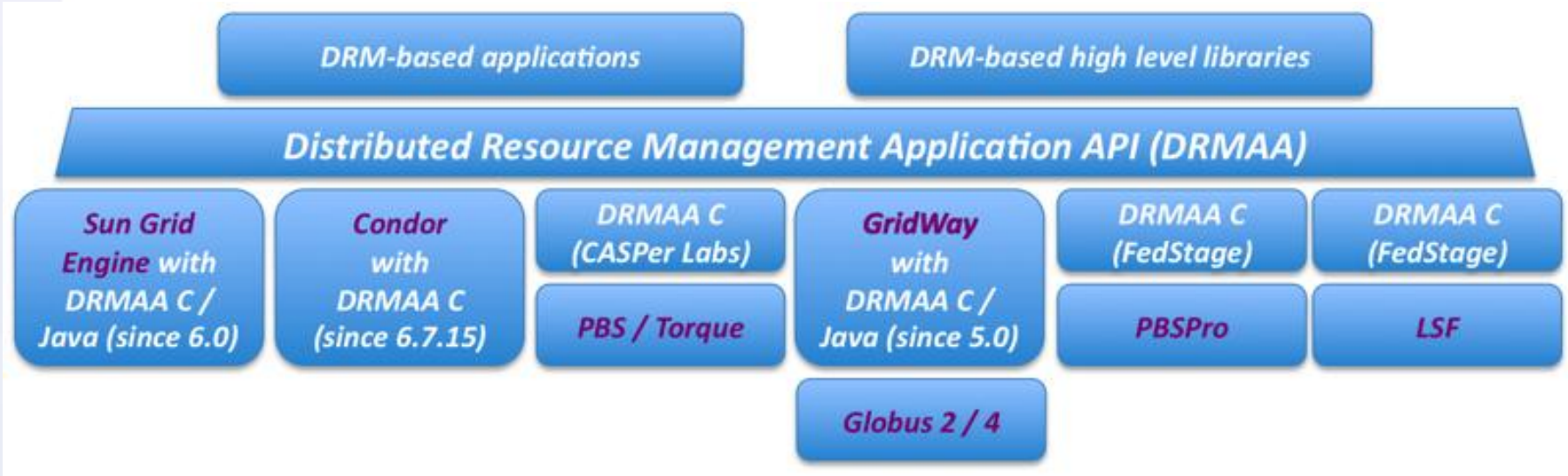
Запуск параллельных программ



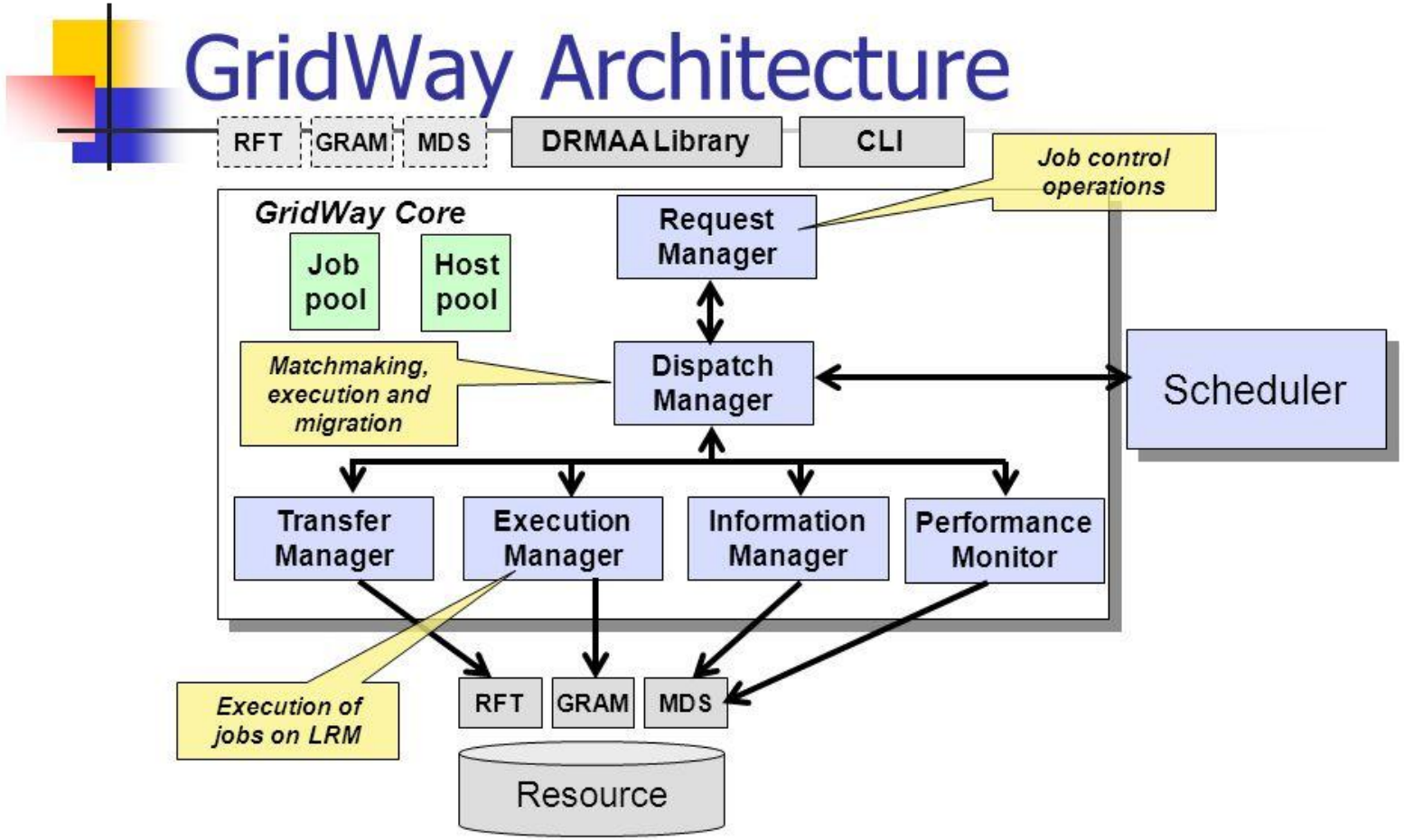




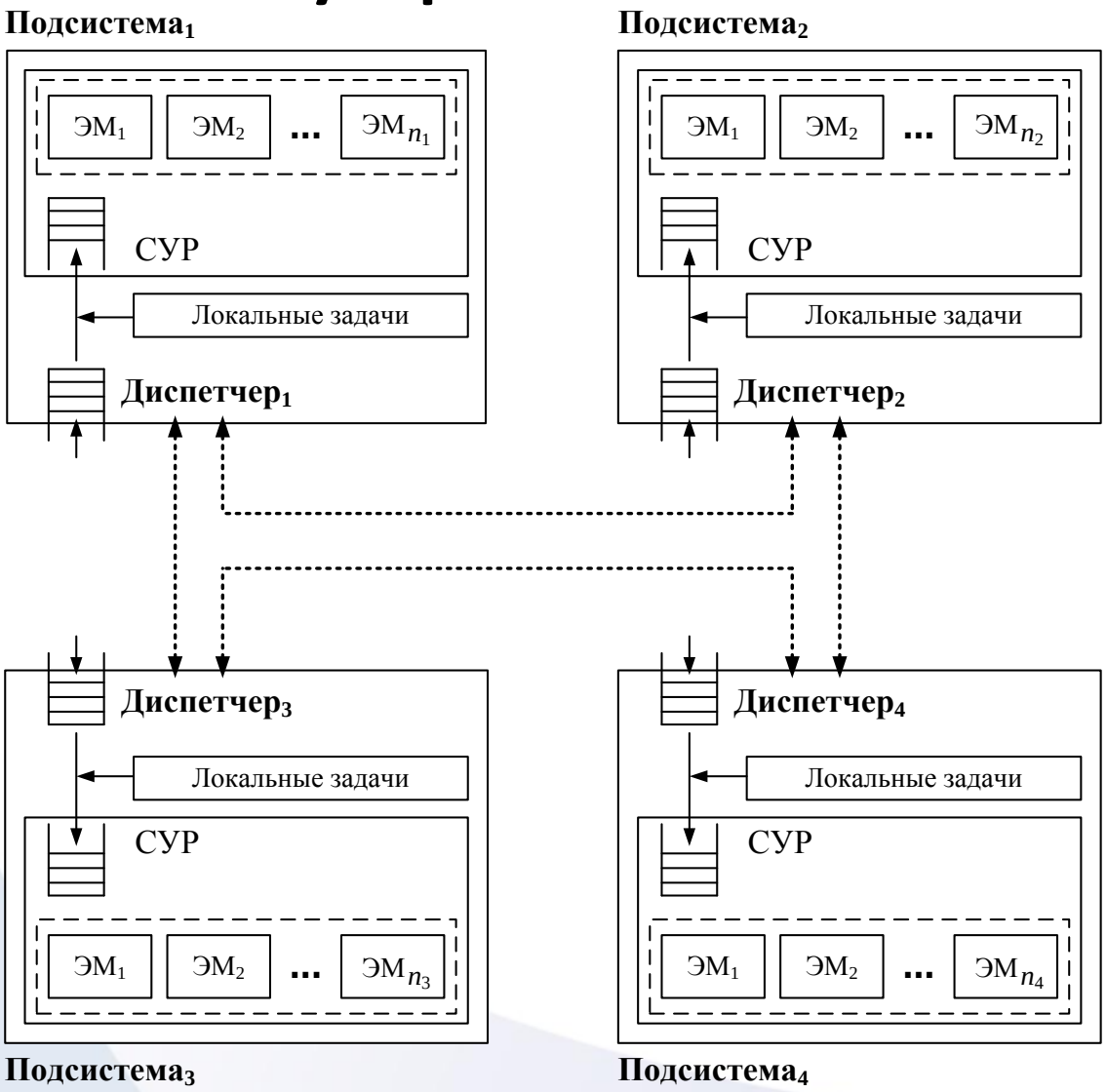
Мультикластерные ВС



GRIDWAY


GridWay Architecture
RFT
GRAM
MDS
DRMAA Library
CLI
Job control operations
GridWay Core
Job pool
Host pool
Request Manager
Matchmaking, execution and migration
Dispatch Manager
Scheduler
Transfer Manager
Execution Manager
Information Manager
Performance Monitor
Execution of jobs on LRM
RFT
GRAM
MDS
Resource

Децентрализованное управление



Литература

Хорошевский В.Г. Архитектура вычислительных систем. Учебное пособие. – М.: МГТУ им. Н.Э. Баумана, 2005; 2-е издание, 2008.

Хорошевский В.Г. Инженерные анализ функционирования вычислительных машин и систем. – М.: “Радио и связь”, 1987.